UNIOSUN JOURNAL OF BASIC AND APPLIED SCIENCES
OSUN STATE UNIVERSITY, OSOGBO
UJBAS

ISSN: 3121-9837
www.ujbs.uniosun.edu.ng/ujbas
ujbas@uniosun.edu.ng

GEO-ADDITIVE REGRESSION MODELLING OF HIGH-DIMENSIONAL DATA WITH COUNT RESPONSE

[1]Saheed O. Jabaru, [2]Kamoru Jimoh
E-mail: saheedjabar@alhikmah.edu.ng

Authors Affiliation:

[1, 2]Department of Physical Sciences, Al-Hikmah University, Kwara State, Nigeria

History:

Volume 1, Number 1
Published: 10/04/2026

Keywords:

Count, High-dimensional, Geo-additive, Regression, Poisson, Negative-Binomial

ABSTRACT

High-dimensional datasets with count responses often arise in the field of epidemiology, environmental science, and socio-economic studies. Traditional regression methods often assume linear relationships and overlook spatial dependence, which can result in biased parameter estimates and unreliable inference when complex data structures are present. This study proposes a methodological framework for modelling high-dimensional count data with mixed predictors by integrating penalized variable selection techniques with Bayesian geostatistical regression modelling. Real-life data from the Nigeria Demographic and Health Survey (NDHS-6), as well as simulated data that mimics the real-life data, were employed. Simulated datasets containing 250 predictors were generated for two sample sizes (n = 50 and n = 100) to mimic real-world data scenarios. Four penalty-based methods: Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, Smoothly Clipped Absolute Deviation (SCAD), and Minimax Concave Penalty (MCP) were applied to detect relevant predictors before model estimation. The selected predictors were subsequently incorporated into multiple geo-additive regression models to account for nonlinear and spatial effects. Model performance was evaluated using the Deviance Information Criterion. The results indicate that the performances of the variable selection techniques become comparable as the sample size increases. Among the models for simulated data, Model 3 and Model 14 were identified as optimal for sample sizes 50 and 100, respectively. The estimated smooth functions revealed nonlinear relationships between metrical predictors and the response variable, while spatial effects indicated moderate heterogeneity across Nigerian states. The significant predictors of TCEB are geopolitical zone, previous residence, educational attainment, terminated pregnancy, contraceptive usage, woman's healthcare, and partner's age. Significant regional variations in fertility rates were identified within Nigeria, underscoring the importance of data characteristics in spatial variability. This study contributes to statistical methodology by providing an integrated framework for identifying significant predictors and modelling spatially dependent count responses in high-dimensional settings.

1. INTRODUCTION

Count data arise in many scientific fields where the response variable represents the number of occurrences of an event within a specified period or spatial unit. Such data take non-negative integer values and frequently appear in epidemiology, environmental studies, public health, and socio-economic research. Examples include the number of disease cases recorded within a region, the number of accidents occurring on a road network, or counts of environmental incidents within geographical areas. Because count responses are discrete and often exhibit skewness and overdispersion, classical linear regression models are generally inappropriate for analysing such data. Consequently, statistical models such as the Poisson and Negative Binomial regression models and the response variable through smooth functions

have been widely used for analysing count responses (Cameron & Trivedi, 2013; Hilbe, 2011).

In many modern applications, researchers are confronted with high-dimensional datasets, where the sample size is smaller compared to the number of predictors. This situation often leads to challenges such as multicollinearity, overfitting, and unstable parameter estimates, which can significantly reduce the predictive performance and interpretability of statistical models (Hastie *et al.,* 2009). In such contexts, variable selection techniques become essential in identifying the most relevant predictors and improving model parsimony. Penalized regression methods such as the Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, Smoothly Clipped Absolute Deviation (SCAD), and Minimax Concave Penalty (MCP) have been widely adopted for this purpose because they simultaneously perform coefficient estimation and variable selection (Zou & Hastie, 2005; Desboulets, 2018).

Beyond high dimensionality, many real-world datasets also exhibit nonlinear relationships and spatial dependence. For example, environmental exposures, demographic characteristics, or geographical factors may influence the occurrence of events in a nonlinear manner across different regions. Meanwhile, the general linear model assumes a linear relationship between predictors and the response variable, and therefore, may fail to capture such complex patterns. The Generalized Additive Model (GAM) provides a flexible alternative by allowing nonlinear relationships between predictors (Wood, 2017). When spatial information is incorporated into this framework, the resulting geo-additive model can simultaneously account for nonlinear covariate effects and spatial variation across geographic regions (Kammann & Wand, 2003).

Despite these methodological advances, several limitations remain in the current literature. Previous studies have either focused on variable selection methods for high-dimensional data or on geo-additive modelling of spatial count data. Furthermore, many studies tend to ignore the joint effects of high dimensionality, nonlinear relationships, and spatial heterogeneity, which may lead to biased estimates and an incomplete understanding of spatial processes.

Addressing this gap is important because high-dimensional spatial count data frequently arise in practical applications such as disease mapping, environmental health assessment, agricultural productivity analysis, and socio-economic studies, where identifying relevant predictors and understanding spatial variability are crucial for effective decision-making and policy formulation. Therefore, developing robust modelling frameworks that can simultaneously handle high-dimensional predictors, nonlinear effects, and spatial dependence is essential for improving statistical inference and predictive accuracy in such contexts.

In view of these challenges, this study proposes a methodological framework for geo-additive regression

regression modelling of selected predictors from high-dimensional data with count responses. The study integrates penalized variable selection techniques with Bayesian geo-additive modelling to identify relevant predictors while accounting for nonlinear and spatial effects. Specifically, the objectives of the study are to:

1. Apply penalized regression techniques to select relevant predictors from high-dimensional datasets containing mixed predictors.
2. Develop geo-additive regression models that incorporate nonlinear effects of metrical predictors and spatial effects across geographical regions.
3. Compare the performance of alternative geo-additive models using appropriate model selection criteria.

Combining variable selection procedures with geo-additive regression modelling provides an integrated approach for analysing high-dimensional data with spatial structure. The proposed framework offers insights for identifying influential predictors and understanding spatial variability in complex datasets encountered in fields such as epidemiology, environmental science, and socio-economic research.

2.0 MATERIALS AND METHODS

2.1 Study Region

The region used for data mapping is Nigeria, a country with 36 States and a Federal Capital Territory. It is further subdivided into 744 local government areas.

2.2 Data simulation scheme for predictors

A. Metrical predictors: These are discrete or continuous variables. The discrete data are generated using the Poisson distribution, while the Uniform and Normal distributions are used to generate continuous data, with the parameters stated as follows.

1. **Poisson (λ)**
   $\lambda_{min} = 0.1;\ \lambda_{max} = 4.8, Interval = 0.1$

2. **Uniform (a, b)**
   a: 1 throughout
   $b_{min} = 0.1;\ b_{max} = 4.8, Interval = 5$

3. **Normal: (μ, σ)**
   $\mu_{min} = 1;\ \mu_{max} = 17, Interval = 1$

   $\sigma = 0.5\ throughout$

B. Categorical Predictors: These are of different categories. However, categorical data of the following categories are simulated with the associated probabilities:

1. **Two-factor variable:** $P_1 : P_2 = 0.72 : 0.28$
2. **Three-factor variable:** $P_1 : P_2 : P_3 = 0.42 : 0.28 : 0.30$
3. **Four-factor variable:** $P_1 : P_2 : P_3 : P_4 = 0.13 : 0.18 : 0.26 : 0.43$
4. **Five-factor variable:** $P_1 : P_2 : P_3 : P_4 : P_5 = 0.07 : 0.13 : 0.26 : 0.20 : 0.32$
5. **Six-factor variable:** $P_1 : P_2 : P_3 : P_4 : P_5 : P_6 = 0.17 : 0.10 : 0.13 : 0.21 : 0.19 : 0.20$

The data were simulated for n = (50, 100). From the 250 predictors in each case, with the sparsity assumption, it was assumed that 20 metrical predictors and 30 dummy variables, making 20% of the total predictors, are significant.

Lastly, the response variable was simulated in R with the function:

$$y = rpois(n, \lambda_i) \quad (1)$$

where:

$$\lambda_i = exp(\beta_0 + \sum_{i=1}^{n} \beta_i x_i) \quad (2)$$

2.3 Real-life Data

The real-life data were extracted from the 6th Nigeria Demographic and Health Survey (NDHS) that was conducted in the year 2018 (NDHS-6). NDHS data provides information on health and demographic variables in the country. Specifically, the NDHS gathers information on women's fertility, the use of family planning techniques and their awareness, women's and children's nutritional status, child and maternal health, women's empowerment, childhood and adult mortality, female genital mutilation, domestic violence, malaria prevalence, etc. The 2018 NDHS program covers women aged 15-49 years and men aged 15-59 years in 42,000 selected households from the 36 States and Federal Capital Territory (Abuja).

Assessing the NDHS-6 data requires registration through the DHS Program's website at www.DHSprogram.com/Data. This process ensures that data usage complies with ethical standards and promotes the confidentiality of respondents. The study variable is the Total Children Ever Born (TCEB) by Nigerian women as at the 2018 NDHS survey, while the predictors are presented in Table 1 below.

Table 1: Predictors of TCEB

	Categorical Predictors	
Variables	**Description**	**Categories**
Respondent_Region	The geopolitical zone of the respondent	North-West North-East North-Central South-West South-South South East
Residence_type	Type of place of residence	Urban Rural
Previous_residence	Type of place of previous residence	City Town Countryside
Previous_state	State of previous residence	36 States of Nigeria + FCT
Education_level	The highest educational level of the respondent	No education Primary

		Secondary Higher
Wealth_index	Wealth index	Poorest Poorer Middle Richer Richest
Had_terminated_pregnancy	Whether the respondent ever had a terminated pregnancy	No Yes
Current_contraceptive	Current contraceptive method used by the respondent	No method Folkloric method Traditional method Modern method
Last_birth_CS	The last birth was through CS	No Yes
Anaemia_level	Anaemia level in the body of the respondent	Severe Moderate Mild Not anaemic
Marital_status	Current marital status of the respondent	Never in union Married Living with a partner Widowed Divorced Separated
Respondent_healthcare	The person who decides on the respondent's health care needs	Respondent alone Respondent and partner Partner alone Someone else Others

Metrical Predictors	
Respondent_Age	Age of the respondent
Number_in_household	Number of household members
Under_five_children	Number of children who are 5 years and below in the household
Births_in_last_five_yrs	Number of births in the last five years
Age_at_first_birth	Age of respondent at first birth
Age_first_cohabitation	Age of respondent at first cohabitation
Years_first_cohabitation	Years since the first cohabitation
Age_first_sex	Age of respondent at first sex
Time_since_last_sex	Time since last sex
Months_of_abstinence	Months of abstinence from sex by the respondent
Age_of_partner	Husband/Partner's age
Preceding_birth_interval	Preceding birth interval
Succeeding_birth_interval	Succeeding birth interval

A total of ninety-three (p = 93) predictors (including dummy variables) were identified as potential predictors of *total children ever born*(y). Forty (40) and Eighty (80) women with complete information were randomly selected without replacement.

2.4 Variable Selection Techniques

Four penalty-based criteria were used to select significant predictors used in the geo-additive modelling (2.5). These are: Least Absolute Shrinkage Selection Operator (LASSO), Elastic Net with strengths α = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9; Smoothly Clipped Absolute Deviation (SCAD) and Minimax Concave Penalty (MCP) with strengths γ = 3, 3.7,

4, 5, 10, 15, 20, 30, 50, and 100. The performances of the variable selection criteria were compared at 4, 5, and 10 cross-validation folds.

According to Desboulets (2018), a parsimonious set of predictors can be achieved through the test-based, screening-based, or penalty-based procedures. However, the penalty-based procedure was used in this study due to the limitations of the other two procedures, among which are:

1. inconsistent variable selection of the test-based procedures because of the addition and subtraction of predictors.
2. screening-based procedure only ranks variables but does not select them.

2.5 Geo-additive Regression Model Formulation

Given:

$v_i = (v_{i1}, \cdots, v_{iq})$ as a vector of q categorical predictors selected in 2.4;

$x_{ij} = (x_{i1}, \cdots x_{ip})$ as the vector of metrical predictors selected in 2.4; and

$\phi_k (k = 1, \ldots, 37)$ as the spatial variable from the Nigeria map object;

The geo-additive regression model for the data in this study is given by:

$$ln(\lambda_i) = \beta_0 + v_i'\beta_i + \sum_{j=1}^{p} f_j(x_{ij}) + f_{spat}(\phi_k) \quad (3)$$

where:

β_0 = model intercept;

β_i = vector of parameters for v_i;

f_j = smooth functions of x_{ij}; and

f_{spat} = smooth functions of ϕ_k

2.6 Prior and Posterior distributions for the geo-additive regression parameters

For geo-additive regression modelling with the Bayesian approach, as suggested by Fahrmeir et al. (2013), the vectors of regression coefficients are considered random variables to incorporate uncertainty into the model, while appropriate prior distributions are assigned to the predictors. An independent flat prior was assumed for the categorical predictors to assign them equal weights, while smoothness of the metrical predictors was achieved through a first-order random walk model to control their wiggliness and avoid overfitting (Brezger & Lang, 2006).

2.7 Model Selection

The best geo-additive regression model was selected through the Deviance Information Criterion (DIC) since full Bayesian models with MCMC were fitted. It is given by:

$$DIC = \overline{D} + p_D \quad (4)$$

where:

$\overline{D}$ = posterior mean deviance

p_D = effective number of parameters

When full MCMC output is not available, the DIC value is approximated as:

$$DIC = -2\log L(y/\theta) + 2p_D$$

where:
$L(y/\theta)$ = likelihood of the data
θ = model parameters

The model with the minimum DIC is the best.

2.8 Software used for data analyses

2.8.1 Data Simulation

All simulations are done using R package version 4.1.0, obtained from the Comprehensive R Archive Network.

2.8.2 Variable selection

This was achieved in R through these two add-on packages:

(1) "glmnet" package version 4.1-2: Here, the regularization strength isgiven by:

$$Regularization\ strength = \begin{cases} 1; & for\ LASSO \\ 0 < \alpha < 1; & for\ Elastic\ Net \end{cases}$$

(2) "ncvreg" package version 3.13.0: Here, the regularization strength is given by:

$$Regularization\ strength = \begin{cases} \gamma > 2; & for\ SCAD \\ \gamma > 1; & for\ MCP \end{cases}$$

2.8.3 Geo-additive regression model

This was also achieved in R through the "R2BayesX" package version 0.3-1.

3.0 RESULTS

The results of simulated data are presented first, followed by real-life data.

Table 2: Results of variable selection for simulated data

Method	*Sample size*	*CV Folds*	*Predictors selected*	*Accuracy (%)*
LASSO/Elastic Net	n = 50	4 - 10	21 - 38	73.6 - 76.4
	n = 100	4 - 10	7 - 25	78.8 - 79.6
SCAD	n = 50	4 - 10	14 - 17	77.2 - 78.4
	n = 100	4 - 10	7 - 9	78.8 - 79.6
MCP	n = 50	4 - 10	0 - 17	0.0 - 80.4
	n = 100	4 - 10	6 - 9	78.8 - 79.6

As presented in Table 2, the number of selected predictors from the penalized regressions ranges from 6 to 25 for n = 100, while it ranges from 0 to 38 for n = 50.

Table 3: Simulated data description for geo-additive models

Model	*n*	Predictors			Mean (*y*)	Var (*y*)	Probability Distribution used
		Metrical	Categorical	Total (p)			
1S	*100*	*4*	*2*	*6*	1.43	1.32	Poisson
2S	*100*	*4*	*3*	*7*	1.50	1.83	Negative Binomial
3S	*100*	*4*	*4*	*8*	1.70	2.01	Negative Binomial
4S	*100*	*4*	*5*	*9*	1.57	1.54	Poisson
5S	*100*	*7*	*5*	*12*	1.56	1.62	Negative Binomial
6S	*100*	*12*	*12*	*24*	1.59	1.74	Negative Binomial
7S	*50*	*4*	*3*	*7*	1.48	1.68	Negative Binomial
8S	*50*	*6*	*3*	*9*	1.58	1.84	Negative Binomial
9S	*50*	*5*	*5*	*10*	1.50	1.72	Negative Binomial
10S	*50*	*5*	*6*	*11*	1.42	1.60	Negative Binomial
11S	*50*	*6*	*7*	*13*	1.62	1.46	Poisson
12S	*50*	*6*	*7*	*13*	1.58	1.31	Poisson
13S	*50*	*6*	*8*	*14*	1.60	1.92	Negative Binomial
14S	*50*	*7*	*9*	*16*	2.04	2.41	Negative Binomial
15S	*50*	*6*	*11*	*17*	2.04	2.41	Negative Binomial

Table 3 reveals the predictor types and the probability distribution used for geo-additive regression modelling.

The full functional form of the geo-additive models for the simulated data is as follows:

Model 1S:

$$\eta_1 = \beta_0 + \beta_1 x_{111} + \beta_2 x_{117} + f_1(x_{15}) + f_2(x_{18}) + f_3(x_{43}) + f_4(x_{75}) + f_{spat1}(States) \quad (5)$$

Model 2S:

$$\eta_2 = \beta_0 + \beta_1 x_{106} + \beta_2 x_{111} + \beta_3 x_{117} + f_1(x_{15}) + f_2(x_{18}) + f_3(x_{43}) + f_4(x_{75}) + f_{spat2}(States) \quad (6)$$

Model 3S:

$$\eta_3 = \beta_0 + \beta_1 x_{106} + \beta_2 x_{111} + \beta_3 x_{114} + \beta_4 x_{117} + f_1(x_{15}) + f_2(x_{18}) + f_3(x_{43}) + f_4(x_{75}) + f_{spat3}(States) \quad (7)$$

Model 4S:

$$\eta_4 = \beta_0 + \beta_1 x_{106} + \beta_2 x_{111} + \beta_3 x_{114} + \beta_4 x_{117} + \beta_5 x_{150} + f_1(x_{15}) + f_2(x_{18}) + f_3(x_{43}) + f_4(x_{75}) + f_{spat4}(States) \quad (8)$$

Model 5S:

$$\eta_5 = \beta_0 + \beta_1 x_{106} + \beta_2 x_{111} + \beta_3 x_{114} + \beta_4 x_{117} + \beta_5 x_{150} + f_1(x_{15}) + f_2(x_{18}) + f_3(x_{27}) + f_4(x_{43}) + f_5(x_{75}) + f_6(x_{85}) + f_7(x_{98}) + f_{spat5}(States) \quad (9)$$

Model 6S:

$$\eta_6 = \beta_0 + \beta_1 x_{106} + \beta_2 x_{110} + \beta_3 x_{111} + \beta_4 x_{114} + \beta_5 x_{117} + \beta_6 x_{118} + \beta_7 x_{121} + \beta_8 x_{123} + \beta_9 x_{125} + \beta_{10} x_{138} + \beta_{11} x_{144} + \beta_{12} x_{150} + f_1(x_7) + f_2(x_{15}) + f_3(x_{18}) + f_4(x_{21}) + f_5(x_{27}) + f_6(x_{31}) + f_7(x_{37}) + f_8(x_{43}) + f_9(x_{73}) + f_{10}(x_{75}) + f_{11}(x_{85}) + f_{12}(x_{98}) + f_{spat6}(States) \quad (10)$$

Model 7S:

$$\eta_7 = \beta_0 + \beta_1 x_{103} + \beta_2 x_{109} + \beta_3 x_{137} + f_1(x_9) + f_2(x_{18}) + f_3(x_{39}) + f_4(x_{75}) + f_{spat7}(States) \quad (11)$$

Model 8S:

$$\eta_8 = \beta_0 + \beta_1 x_{103} + \beta_2 x_{109} + \beta_3 x_{120} + \beta_4 x_{117} + f_1(x_6) + f_2(x_9) + f_3(x_{18}) + f_4(x_{39}) + f_5(x_{75}) + f_{spat8}(States) \quad (12)$$

Model 9S:

$$\eta_9 = \beta_0 + \beta_1 x_{103} + \beta_2 x_{109} + \beta_3 x_{120} + \beta_4 x_{128} + \beta_5 x_{137} + f_1(x_6) + f_2(x_9) + f_3(x_{18}) + f_4(x_{39}) + f_5(x_{75}) + f_{spat9}(States) \quad (13)$$

Model 10S:

$$\eta_{10} = \beta_0 + \beta_1 x_{103} + \beta_2 x_{109} + \beta_3 x_{120} + \beta_4 x_{128} + \beta_5 x_{137} + \beta_6 x_{140} + f_1(x_6) + f_2(x_9) + f_3(x_{18}) + f_4(x_{39}) + f_5(x_{75}) + f_{spat10}(States) \quad (14)$$

Model 11S:

$$\eta_{11} = \beta_0 + \beta_1 x_{103} + \beta_2 x_{104} + \beta_3 x_{109} + \beta_4 x_{120} + \beta_5 x_{128} + \beta_6 x_{135} + \beta_7 x_{140} + f_1(x_6) + f_2(x_9) + f_3(x_{18}) + f_4(x_{20}) + f_4(x_{39}) + f_5(x_{75}) + f_{spat11}(States) \quad (15)$$

Model 12S:

$$\eta_{12} = \beta_0 + \beta_1 x_{103} + \beta_2 x_{109} + \beta_3 x_{114} + \beta_4 x_{120} + \beta_5 x_{128} + \beta_6 x_{137} + \beta_7 x_{140} + f_1(x_6) + f_2(x_9) + f_3(x_{18}) + f_4(x_{24}) + f_5(x_{39}) + f_6(x_{75}) + f_{spat12}(States) \quad (16)$$

Model 13S:

$$\eta_{13} = \beta_0 + \beta_1 x_{103} + \beta_2 x_{109} + \beta_3 x_{114} + \beta_4 x_{120} + \beta_5 x_{128} + \beta_6 x_{137} + \beta_7 x_{140} + \beta_8 x_{145} + f_1(x_6) + f_2(x_9) + f_2(x_{18}) + f_3(x_{24}) + f_4(x_{39}) + f_5(x_{75}) + f_{spat13}(States) \quad (17)$$

Model 14S:

$$\eta_{14} = \beta_0 + \beta_1 x_{103} + \beta_2 x_{109} + \beta_3 x_{114} + \beta_4 x_{120} + \beta_5 x_{127} + \beta_6 x_{128} + \beta_7 x_{137} + \beta_8 x_{140} + \beta_9 x_{145} + f_1(x_6) + f_2(x_9) + f_3(x_{18}) + f_4(x_{24}) + f_5(x_{39}) + f_6(x_{73}) + f_7(x_{75}) + f_{spat}(States) \quad (18)$$

Model 15S:

$$\eta_{15} = \beta_0 + \beta_1 x_{103} + \beta_2 x_{109} + \beta_3 x_{114} + \beta_4 x_{120} + \beta_5 x_{127} + \beta_6 x_{128} + \beta_7 x_{135} + \beta_8 x_{137} + \beta_9 x_{140} + \beta_{10} x_{145} + \beta_{11} x_{146} + f_1(x_6) + f_2(x_9) + f_3(x_{18}) + f_4(x_{24}) + f_5(x_{39}) + f_6(x_{75}) + f_{spat}(States) \quad (19)$$

Table 4: Comparison of the geo-additive regression models

Model	*N*	*Metrical Predictors*	*Categorical predictors*	*DIC*
Model 1S	100	4	2	184.261
Model 2S	100	4	3	166.051
Model 3S	100	4	4	156.570*
Model 4S	100	4	5	171.632
Model 5S	100	7	5	170.611
Model 6S	100	12	12	160.746
Model 7S	50	4	3	89.463
Model 8S	50	5	3	77.769
Model 9S	50	5	5	89.2334
Model 10S	50	5	6	90.6162
Model 11S	50	6	6	84.6962
Model 12S	50	6	7	86.3474
Model 13S	50	6	8	69.0774
Model 14S	50	7	9	45.0867**
Model 15S	50	6	11	47.592

*Minimum value for n = 100 **Minimum value for n = 50

Table 4 shows the results of the model selection criterion (DIC) for the 15 geo-additive models. It was observed that Model 3S and Model 14S are the optimal models for sample sizes $n = 100$ and $n = 50$, respectively. Hence, their results are fully presented as follows:

Table 5: Parametric coefficients for Model 3S and Model 14S

		Parameter estimates		95% credible interval	
		β	Sd	2.5%	97.5%
Model 3S	(Intercept)	0.0197	0.42219	-0.8347	0.7933
	$x106_0$ (ref)	0			
	$x106_1$	-0.0228	0.1812	-0.3900	0.3252
	$x111_0$ (ref)	0			
	$x111_1$	0.4126*	0.1710	0.0778	0.7320
	$x114_0$ (ref)	0			
	$x114_1$	0.6386	0.4505	-0.2401	1.5628
	$x114_2$	0.3878	0.4199	-0.4125	1.2705
	$x114_3$	0.0913	0.46224	-0.7725	1.0207
	$x114_4$	0.3116	0.4377	-0.5241	1.2164
	$x117_0$ (ref)	0			
	$x117_1$	-0.0158	0.2032	-0.4116	0.3814
	$x117_2$	0.0996	0.1995	-0.2680	0.4803
	(Intercept)	-53.9658*	22.6382	-98.0035	-17.0947
	$x103_0$ (ref)	0			
	$x103_1$	-19.0256*	8.9950	-37.8101	-2.9448
	$x103_2$	-8.2896*	4.8129	-17.8212	-0.3554
	$x103_3$	-7.6569*	4.6144	-17.7522	-0.3650
	$x109_0$ (ref)	0			
	$x109_1$	9.3046	7.1826	-2.7454	26.2599
	$x109_2$	14.1683*	7.7303	4.3303	34.0561
	$x109_3$	27.8468*	14.3096	6.0116	58.5279
	$x109_4$	7.1592*	6.3921	-3.3747	23.3453

Model 14S	$x114_0$ (ref)	0			
	$x114_1$	12.5990*	6.1287	1.6326	22.6351
	$x114_2$	15.5175*	8.3432	2.0904	27.0734
	$x114_3$	19.4004*	9.1778	4.5952	37.8062
	$x114_4$	13.3899*	7.0244	2.0322	24.8601
	$x120_0$ (ref)	0			
	$x120_1$	17.4697*	7.3381	3.6210	30.5136
	$x120_2$	8.7016*	3.5462	2.9732	16.3006
	$x120_3$	7.3433	5.1226	-0.2326	16.4081
	$x120_4$	-4.7044	3.5326	-11.8704	1.8501
	$x120_5$	-2.6730	3.0916	-8.4982	3.2521
	$x127_0$ (ref)	0			
	$x127_1$	6.7255*	2.3475	2.3859	11.4581
	$x127_2$	9.7139*	3.7092	3.2142	15.7143
	$x128_0$ (ref)	0			
	$x128_1$	18.6131*	6.8539	5.5462	29.4135
	$x128_2$	16.6130*	5.5279	6.5927	26.1400
	$x128_3$	14.1887*	5.2273	5.3907	23.6430
	$x137_0$ (ref)	0			
	$x137_1$	5.7732	3.5266	-0.2212	11.0170
	$x137_2$	7.0330*	3.3861	1.5720	13.1701
	$x140_0$ (ref)	0			
	$x140_1$	17.4310*	6.7182	4.4750	29.6732
	x1402	8.3005*	4.1477	0.2383	17.2577
	x1403	9.1534	5.8109	-0.0147	19.8615
	$x140_4$	11.1343*	5.0148	1.6454	18.6173
	$x140_5$	12.9635*	6.3745	2.8066	21.9878
	$x145_0$ (ref)	0			
	$x145_1$	7.2336*	3.5540	2.2384	15.7655
	$x145_2$	0.8989	2.5107	-3.0931	6.4485
	$x145_3$	0.1281	2.0587	-3.8013	3.7014
	$x145_4$	3.3550	2.5909	-1.3745	8.4004
	$x145_5$	2.4265	2.4190	-1.6896	7.1457

*Estimates are significant at 5% significance level. *Ref = Reference Category;* β = *posterior mean*

In Table 5, level 1 of both x_{106} and x_{117} in Model 3S has a lower impact on the response compared to their reference categories, while other factor levels have higher effects on the response variable than their reference categories. Moreover, the effects of levels 1, 2, and 3 of x_{103} and levels 4 and 5 of x_{120} in Model 14S are lower than those of the reference categories, while other factor levels have higher effects.

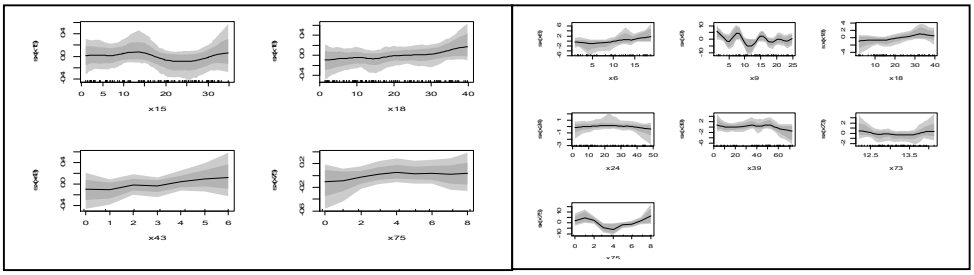

Figure 1: Non-linear effects of metrical predictors in Model 3S and Model 14S

Each metrical predictor for Models 3S and Model 14S in Figure 1 shows a unique non-linear relationship with the response variable.

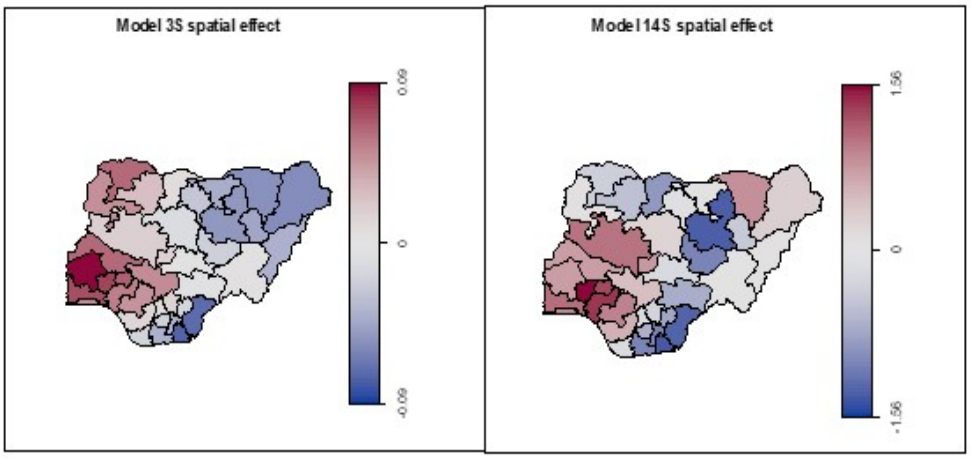

Figure 2: Spatial distributions of response variables within Nigeria for Models 3S and14S

The differences between the highest and the lowest values of response variables on the maps in Figure 2 for Models 3S and 14S are ±0.09 and ±1.56, respectively.

Real-life Data Results

The dimensionality of the data was reduced by considering scenarios with the highest percentage of accuracy in the simulated data variable selection. The result of variable selection for the real-life data are presented in Table 6 below:

Table 6: Number of variables selected for real-life data

Criterion	k-fold	Tuning parameter	Number of predictors selected	
			n = 80, p = 93	*n = 40, p = 93*
LASSO	4	Nil	12	0
	4	0.5	12	0
Elastic Net	5	0.5	0	0
	10	0.9	1	0
	4	3	10	1
SCAD	5	3	0	6
	10	3	1	0
	4	10	12	4
MCP	5	3	0	0
	10	3	0	0

The 12 features selected in Table 6 are the predictors used for Model 1R, 10 features for Model 2R, 6 features for Model 3R, and 4 features for Model 4R. The results of the two best Models (1R and 2R) are presented below.

Table 7: Parametric coefficients of Models 1R and 2R

	Parameter estimates				95% credible interval			
	Model 1R		**Model 2R**		**Model 1R**		**Model 2R**	
	β	Sd	β	Sd	Lower	Upper	Lower	Upper
Intercept	0.87	0.585	0.830	0.614	-0.469	1.840	-0.323	1.999
Region (*Ref: North-Central*)	0		0					
North-East	0.051	0.253	0.084	0.238	-0.421	0.604	-0.363	0.556
North-West	-0.025	0.228	0.037	0.216	-0.487	0.431	-0.367	0.460
South-East	0.394	0.548	0.057	0.440	-0.730	1.424	-0.808	0.885
South-South	-0.055	0.379	-0.133	0.360	-0.826	0.698	-0.851	0.581
South-West	-0.224	0.274	-0.145	0.264	-0.802	0.298	-0.647	0.344
Previous State (*Ref: Abia*)	0		0					
Bauchi	-0.118	0.517	-0.174	0.523	-1.059	0.913	-1.160	0.825
Benue	-0.027	0.475	-0.026	0.488	-0.922	0.944	-0.908	0.978
Borno	-1.565*	0.676	-1.617*	0.683	-2.960	-0.311	-3.023	-0.427
Cross-River	0.161	0.829	0.410	0.817	-1.398	1.847	-1.271	2.074
Ebonyi	-0.815	0.743	-0.579	0.640	-2.212	0.666	-1.789	0.721
Enugu	-1.609*	0.819	-1.262	0.738	-3.180	-0.041	-2.716	0.181
Gombe	-0.086	0.472	-0.211	0.496	-0.948	0.912	-1.099	0.838
Kaduna	-0.321	0.559	-0.616	0.601	-1.404	0.783	-1.720	0.580
Katsina	-1.283	0.751	-1.398	0.753	-2.835	0.063	-3.025	0.028
Kwara	0.071	0.717	-0.386	0.623	-1.241	1.519	-1.542	0.803
Lagos	1.891	11.481	20.815*	7.346	-17.972	18.986	7.396	35.348
Nasarawa	-0.549	0.746	-0.563	0.814	-2.039	0.919	-2.214	1.009
Niger	-0.908	0.745	-0.945	0.727	-2.384	0.540	-2.411	0.436
Ondo	-0.582	0.595	-0.667	0.592	-1.749	0.718	-1.779	0.500
Osun	-0.787	0.806	-0.929	0.890	-2.500	0.686	-2.402	0.862
Outside Nigeria	-0.273	0.606	-0.306	0.582	-1.438	0.999	-1.438	0.899
Sokoto	-0.432	0.583	-0.509	0.651	-1.631	0.634	-1.563	0.720
Taraba	-0.302	0.433	-0.394	0.468	-1.089	0.608	-1.179	0.541
Education Level (Ref: Higher)	0		0					
No education	0.965*	0.411	1.182*	0.378	0.205	1.794	0.455	1.893
Primary	1.174*	0.424	1.367*	0.376	0.392	2.063	0.593	2.082
Secondary	1.006*	0.387	1.130*	0.362	0.268	1.778	0.392	1.768
Had a terminated pregnancy (Ref: No)	0		0					
Yes	-0.070	0.157	NA	NA	-0.372	0.222	NA	NA
Contraceptive (*Ref: Diaphragm*)	0		0					
Injections	0.353	0.427	0.131	0.399	-0.460	1.179	-0.533	0.965
Lactational amenorrhea	-1.746	1.150	-2.264	1.699	-4.394	0.086	-6.168	0.011
Male condom	0.170	0.578	-0.105	0.547	-0.983	1.265	-1.182	1.079
Not using	0.282	0.312	0.076	0.274	-0.326	0.877	-0.418	0.680
Pill	-2.139	11.524	-21.702*	7.404	-18.916	18.206	-35.857	-8.282
Anaemia level (*Ref: Mild*)	0		0					
Moderate	0.093	0.154	0.129	0.154	-0.206	0.394	-0.185	0.420
Not anaemic	-0.028	0.237	-0.042	0.225	-0.489	0.437	-0.502	0.408
Severe	0.648	0.356	0.630	0.334	-0.121	1.326	-0.058	1.275
Respondent healthcare (Ref: husband/partner alone)	0		0					
Respondent alone	-0.435	0.333	NA	NA	-1.169	0.174	NA	NA
Respondent and husband	-0.093	0.146	NA	NA	-0.390	0.202	NA	NA

*Estimates that are significant at 5% significance level. NA = Not Applicable

Table 7 presents the posterior means of the categorical predictors for Models 1R and 2R. Considering those that are significant at 5% significance level in Model 1R, women who previously lived in Abia State have higher TCEB than those who previously lived in Borno and Enugu States. Similarly, in Model 2R, women who previously lived in Abia State have

higher TCEB than those who previously lived in Borno State, while the reverse is the case for women who previously lived in Lagos State.

For both Models, women with a higher level of education have fewer children ever born than those with primary, secondary, or no education. Lastly, in Model 2R, women who are planning their family through *pills* have fewer children ever born than those using a *diaphragm*.

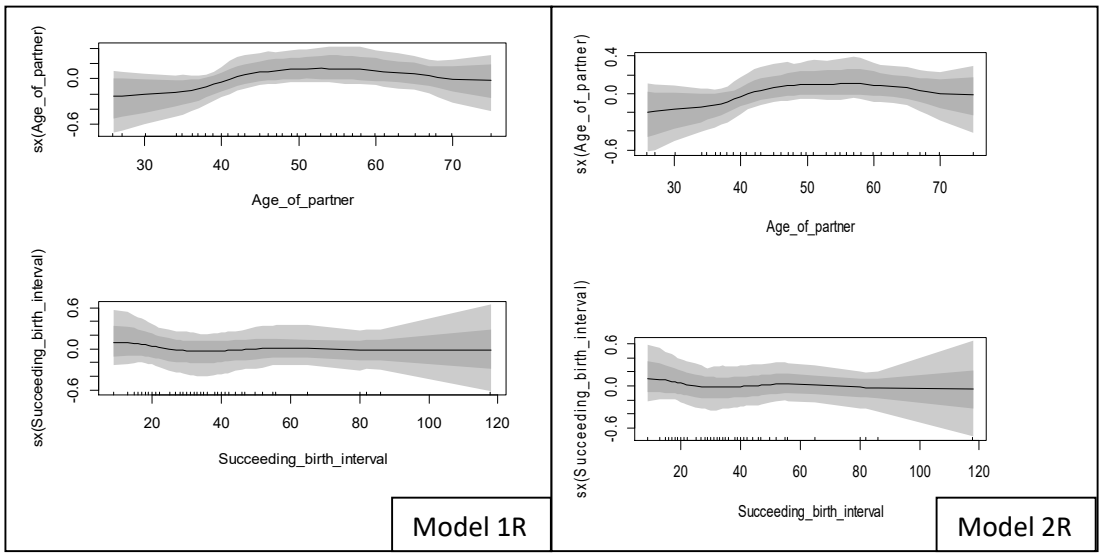

Figure 3: Non-linear effects of metrical predictors (*succeeding birth interval* and *age of partner* in (a) Model 1R (left panel) and (b) Model 2R (Right panel)

In Figure 3, TCEB for partners' age in both panels keeps increasing until 60 years, when it starts declining. Meanwhile, the TCEB decreases with an increase in birth intervals.

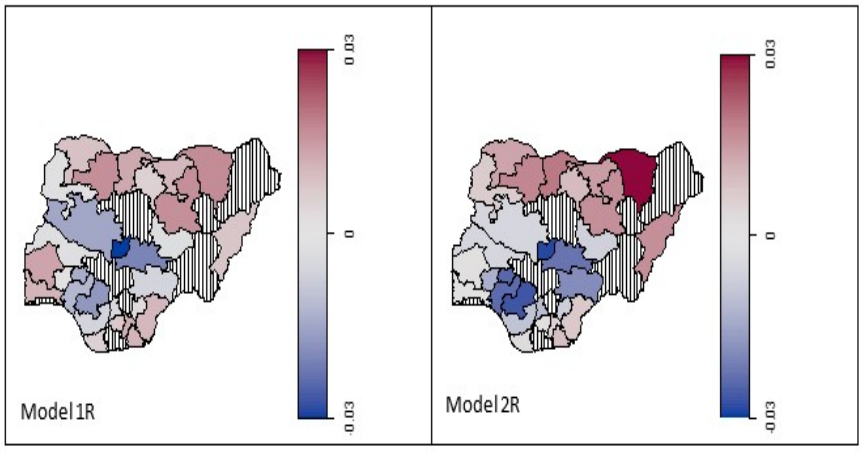

Figure 4: Spatial distribution of total children ever born in Nigeria for (a) Model 1R and Model 2R

In Figure 4, the spatial map of Model 1R shows that the number of children ever born by women in Zamfara, Katsina, Yobe, Bauchi, Oyo, Ogun, Cross River, Adamawa, Akwa-Ibom, and Abia States (red region) is higher than those of other States. Similarly, the spatial map of Model 2R shows that the number of children ever born by women in most North-eastern and North-western States (red region) is higher than that of other States in the country. States with black and white stripes indicate insignificant spatial effect on the response variable.

4.0 DISCUSSION

Results of variable selection revealed that at $n = 50$, MCP has the highest percentage of accuracy (80.4%), while other selection criteria have the same highest value. At $n = 100$, the range of the percentage of accuracy is the same for all the selection criteria. This finding aligns with that of Zhang et al. (2014), who compared penalty-based criteria of variable selection for simulated data with ten-fold cross-validation and asserted that the performances of the selection criteria are not significantly different. However, the result of variable selection in this study contradicts that of Lima et al. (2020), who compared penalized variable selection techniques using low-dimensional datasets with normally distributed responses and concluded that the number of variables selected by each method differs.

The probability distributions of the response variables for 50 and 100 samples agree with previous simulation-based work on count data in environmental and epidemiological studies. The use of the Poisson distribution for data that are under-dispersed or equi-dispersed, and the use of the Negative Binomial distribution for data that are over-dispersed, is justified because the wrong probability distribution will lead to biased estimates (Cameron & Trivedi, 2013; Hilbe, 2011).

Models with a lower number of predictors perform better than those with a higher number of predictors in terms of their DIC values. This implies that geo-additive models with high-dimensional count responses are sensitive to sample size, which supports the conclusion of Hastie et al. (2009) and Wood (2017).

Some factor levels have a positive impact on the response variable while others have a negative impact. Meanwhile, the best model from the larger sample size (Model 3) has one of its factor levels statistically significant to the response variable at 95% credible interval after accounting for spatial and nonlinear effects of metrical predictors. However, the majority of the factor levels from the best model for the smaller sample size (Model 14) contribute significantly to the response variable at 95% credible interval. This aligns with the submission of Gelman *et al.* (2014), where a model with a smaller sample is preferred over the larger sample model.

The curvilinear shapes of all the metrical predictors justify their prior nonlinear assumption as proposed by Wood (2017). The spatial maps revealed variability of the response variables across the

Nigerian States. The spatial range of ±0.09 and ±1.56 for Model 3S and 14S, respectively, indicates moderate spatial heterogeneity, which confirms the presence of unobserved spatial factors, in line with the findings of Kammann & Wand (2003) and Rue *et al.* (2009).

This study's findings regarding predictors of women's fertility rate in Nigeria present several notable parallels and contrasts with existing literature on fertility determinants in sub-Saharan Africa and other developing regions. The study indicates significant regional variations in fertility rate within Nigeria, with higher fertility rates in the North-East compared to North-Central, while there are lower fertility rates in the North-West, South-East, South-West, and South-South regions compared to the North-Central region (Model 1R). For Model 2R, there are lower fertility rates in the South-West and South-South regions compared to the North-Central region. This aligns with findings by Mberu and Reed (2014), who highlighted similar regional fertility disparities, attributing them to differences in socio-economic conditions, cultural norms, and access to healthcare services. Additionally, Garenne (2008) noted that fertility rates are often higher in rural and less developed regions, which often correspond to the northern areas of Nigeria identified in this study.

The significant geographical variations in fertility rates are also supported by the work of Guilmoto (2009), who demonstrated that regional disparities in socio-economic development and cultural practices lead to varied fertility patterns across different areas.

The identification of states with high and low fertility rates in this study echoes the findings of previous regional fertility studies in Nigeria, which also highlighted similar spatial differences (Adebowale, 2019).

The inverse relationship between educational attainment and fertility observed in this study is consistent with the findings of Mashood et al. (2022), Alaba *et al.* (2017), and Amusa & Yahya (2019). Pradhan (2015) also argued that female education played a greater role in delaying the age at which a woman gets married; hence, the consequence of this is that such women will have fewer children before getting to menopausal age. The use of contraceptives was also found to reduce the number of children ever born, which supports the findings of Rahman *et al.* (2022). Age at first birth and level of education also contribute to the lowering of the number of children born.

Furthermore, a study by Gyimah et al. (2012) found that higher levels of female education significantly reduce fertility rates due to delayed marriage and increased use of contraception. Similarly, Caldwell and Caldwell (1987) emphasized that education empowers women with better knowledge and resources for family planning, leading to lower fertility rates. The study's finding that various contraceptive methods reduce fertility rate aligns with Gebre's (2024) study, which concludes that the usage of modern contraceptives by women of reproductive age led to a reduction in fertility rate. It also aligns with Bongaarts and Westoff's (2000) research, which

demonstrated the effectiveness of modern contraceptive methods in lowering fertility rates across different contexts. The significant impact of female sterilization, injections, and male condoms on reducing fertility in this study is in line with previous findings by Cleland et al. (2006), who reported similar trends in family planning studies globally.

Additionally, the association between pregnancy termination and lower fertility contradicts the study by Ibeji *et al.* (2020), which shows a positive correlation between total children born and having terminated a pregnancy. The study's findings on the relationship between the age of the partner and fertility are consistent with existing literature. It also agrees with the findings of Cherie *et al*. (2023), whereby there is a 2.4% increase in children ever born for every unit increase in respondent age. The spatial maps revealed that the number of children born in the northeastern states is higher than that in the northwestern States. Conversely, the number of children born in the North-central, South-west, South-south, and South-eastern states is lower than that in the North-western States (Mashood *et al*., 2022).

In real-life data, the use of cross-sectional data might not allow causal inference to be made. In the simulated data, significant changes in the parameters of the data simulation scheme, sample sizes, and cross-validation folds might lead to different results. However, this study's potential application includes, among others, the identification of high or low-risk/benefit regions of a given phenomenon. This is crucial for understanding disease distribution, environmental health risks, and other spatially dependent phenomena in epidemiology, agriculture, and social sciences.

5.0 CONCLUSION

The results in this study revealed that the performances of the four penalized variable selection criteria are not significantly different at a higher sample size. However, the accuracy of a particular variable selection criterion is still data-dependent. Moreover, an increase in sample size led to the selection of a sparser number of predictors.

The descriptive properties of the simulated count response in this study align with established properties of count data, which justifies the usage of Poisson/Negative-Binomial model for under-dispersed/equi-dispersed response, respectively.

This study also concludes that geo-additive models of predictors from high-dimensional data are sensitive to sample sizes. Models' comparison using DIC revealed that models with fewer predictors perform better than those with a larger number of predictors.

The nonlinear relationship affirmed that a linear model would have resulted in an erroneous estimate. Similarly, the spatial effect revealed significant heterogeneity in the response variable across Nigerian States, which could have been ignored in a generalized additive model.

This study's identification of various predictors of fertility rate in Nigeria provides a comprehensive understanding that aligns with and expands upon existing literature. By incorporating both traditional

and novel predictors, this research offers valuable insights into the multifaceted determinants of fertility rates, emphasizing the need for region-specific and culturally sensitive family planning interventions. The strong influence of educational attainment affirmed the significant role of female education in reducing fertility rates.

Additionally, the significant impact of contraceptive use on fertility rates supports ongoing efforts to increase access to education about family planning methods. The findings regarding marital status and healthcare decision-making dynamics call for strategies that empower women and promote gender equality in decision-making processes. Finally, the spatial and temporal effects identified in this study underline the necessity of continuous monitoring and adaptation of family planning programs to address changing demographic patterns and regional variations. Overall, this study contributes to a deeper understanding of fertility determinants in Nigeria, offering practical recommendations for targeted interventions that can effectively address high fertility rates and promote sustainable population growth.

This study's major contribution to knowledge is the spatial analysis of count response variables using high-dimensional data. It offers a methodological approach for determining the spatial concentration of count responses in high-dimensional datasets with mixed predictors.

6.0 Acknowledgements

The authors acknowledge the National Population Commission and ICF International for providing access to the 2018 Nigeria Demographic and Health Survey (NDHS) data. We also thank Al-Hikmah University for the computational resources provided for this study.

References

Adebowale, A. S. (2019). Ethnic disparities in fertility and its determinants in Nigeria. *Fertility research and practice*, 5(1), 1-16.

Alaba, O. O., Olubusoye, O. E. & Olaomi, J. O. (2017). Spatial patterns and determinants of fertility levels among women of childbearing age in Nigeria. *South African Family Practice*, 59(4), 143-147.

Amusa, L. & Yahya, W. (2019). Stepwise Geo-additive Modelling of the Ideal Family Size in Nigeria. *Turkiye Klinikleri Journal of Biostatistics*, 11(2), 123-132.

Bongaarts, J. & Westoff, C. F. (2000). The potential role of contraception in reducing abortion. *Studies in family planning*, 31(3), 193-202.

Brezger, A. & Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis*, 50(4), 967-991.

Caldwell, J. C. & Caldwell, P. (1987). The Cultural Context of High Fertility in Sub-Saharan Africa. Population and Development Review, 13(3), 409-437.

Cameron, A. C. & Trivedi, P. K. (2013). *Regression analysis of count data,* Cambridge University Press, Cambridge, 566 p

Cherie, N., Getacher, L., Belay, A., Gultie, T., Mekuria, A., Sileshi, S., & Degu, G. (2023). Modelling on the number of children ever born and its determinants among married women of reproductive age in Ethiopia: A Poisson regression analysis. *Heliyon*, 9(3), 1-10.

Cleland, J., Bernstein, S., Ezeh, A., Faundes, A., Glasier, A. & Innis, J. (2006). Family Planning: The Unfinished Agenda. Lancet, 368(9549), 1810-1827.

Desboulets, L. D. D. (2018). A review of variable selection in regression analysis. *Econometrics*, 6(4), 1-16.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. D. (2013). *Regression: Models, Methods and Applications*. Springer Science & Business Media, 698p.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76(376), 817-823.

Garenne, M. (2008). Situations of fertility stall in sub-Saharan Africa. *African Population Studies*, 23(2), 173-188.

Gebre, M. N. (2024). Number of children ever-born and its associated factors among currently married Ethiopian women: evidence from the 2019 EMDHS using negative binomial regression. *BMC Women's Health*, 24(1), 1-11.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. Chapman & Hall/CRC, 675 p.

Guilmoto, C. Z. (2009). The sex ratio transition in Asia. *Population and Development Review*, 35(3), 519-549.

Gyimah, S. O., Adjei, J. K., & Takyi, B. K. (2012). Religion, contraception, and method choice of married women in Ghana. *Journal of religion and health*, 51*(4)*, 1359-1374.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*.). New York, 745 p.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press, Cambridge, 553 p.

Ibeji, J. U., Zewotir, T., North, D., & Amusa, L. (2020). Modelling fertility levels in Nigeria using Generalized Poisson regression-based approach. *Scientific African*, 9(e00494), 1-12.

Kammann, E. E., & Wand, M. P. (2003). Geo-additive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1), 1-18.

Lima, E., Davies, P., Kaler, J., Lovatt, F. & Green, M. (2020). Variable selection for inferential models with relatively high-dimensional data: Between method heterogeneity and covariate stability as adjuncts to robust selection. *Scientific Reports*, 10(1), 1-11.

Mashood, L. O., Ani, C. I., Balogun, O. S. & Abdulazeez, S. A. (2022). A Geo-additive Model of Fertility Level on Female Education among Women of Childbearing Age in Nigeria, *Proceedings of the 1st Faculty of Science International Conference FSIC(2022),* 175-188.

Mberu, B. U., & Reed, H. E. (2014). Understanding subgroup fertility differentials in Nigeria. *Population review*, 53(2), 23-46.

Pradhan, E. 2015. Female education and childbearing: A closer look at the data. World Bank Blog. http://blogs. World Bank. org/health/femaleeducation-and-childbearing-closer-look-data.

Rahman, A., Hossain, Z., Rahman, M. L., & Kabir, E. (2022). Determinants of children ever born among ever-married women in Bangladesh: evidence from the Demographic and Health Survey 2017-2018. *BMJ open*, 12(6), 1-11.

Wood, S. N. (2017). P-splines with derivative-based penalties and tensor product smoothing of unevenly distributed data. *Statistics and Computing*, 27(4), 985-989.

Zhang, K., Yin, F., & Xiong, S. (2014). Comparisons of penalized least squares methods by simulations. *arXiv preprint arXiv:*1405.1796*, 1-9*.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2), 301-320.